Improving Depth Maps by Nonlinear Diffusion

Jianfeng Yin
Centre for Intelligent Machines
McGill University
3480 University Street
Montreal, QC H3A 2A7 Canada

jfyin@cim.mcgill.ca

Jeremy R. Cooperstock
Centre for Intelligent Machines
McGill University
3480 University Street
Montreal, QC H3A 2A7 Canada

jer@cim.mcgill.ca

ABSTRACT

Dense depth maps, typically produced by stereo algorithms, are essential for various computer vision applications. For general configurations in which the cameras are not necessarily parallel or close together, it often proves difficult to obtain reasonable results for complex scenes, in particular in occluded or textureless regions. To improve the depth map in such regions, we propose a post-processing method and illustrate its benefits to applications such as 3D reconstruction or foreground segmentation of a user in a scene.

Keywords
Stereo Matching, Nonlinear Diffusion, 3D Reconstruction, Background Removal.

1 INTRODUCTION

Stereo matching algorithms are an extensively studied topic in computer vision. Dense depth maps produced by these algorithms constitute the basis for many applications including 3D reconstruction, image-based rendering, and scene segmentation. Problems arise, however, in regions of occlusion or lacking texture. Using smoothness constraints, many stereo algorithms treat such regions as having the same or similar depth as neighboring areas, often causing objects to appear larger or wider, an obviously undesirable effect. Various algorithms have been developed to address this problem. For example, Kanade and Okutomi [Kan94, Oku93] use adaptive windows or multiple cameras, and Scharstein and Szeliski [Sch96] aggregate support in stereo matching by nonlinear diffusion instead of using an explicit window. Belhumeur and Mumford[Bel92], Intille and Bobick [Int94], and Geiger *et al*.[Gei92] incorporate occlusion information directly into their stereo matching algorithms by Bayesian methods and use dynamic programming for optimization. Gamble *et al* [Gam87] integrate discontinuity information and depth information by a coupled Markov random field model. More elegant methods adopt a global strategy and use graph cuts to minimize the energy, which can consider such situations explicitly [Boy01, Kol01, Roy98]. A more detailed analysis of many stereo algorithms can be found in Scharstein and Szeliski's paper [Sch02]. Although these approaches solve the problem to a certain extent, they are insufficient to address the complexities of general scenes, such as that illustrated in Fig. 2. Further, while many algorithms consider the binocular case, they cannot be extended to $N(\geq 3)$ camera problems, employing a generic configuration. For such cases, rectification may be difficult, if not impossible, and the range of disparities can be very large.

Improvements to the depth map can be obtained through filtering or interpolation. For example, median filters or morphological filters can fill small gaps and correct depth errors (e.g. [Schm02]), but their ability to do so is rather limited. Linear interpolation techniques (e.g. [Kau01]) can fill gaps along epipolar lines or scanlines when images are rectified. The drawback is that these methods use only the information along one line, which may be difficult to estimate correctly when the epipolar geometry is complicated. Furthermore, such interpolation methods do not con-

WSCG SHORT Communication papers proceedings
WSCG'2004, February 2-6, 2004, Plzen, Czech Republic.

sider the information provided by other neighboring areas. Notwithstanding these efforts, we suggest that further improvements to the depth map may be obtained through post-processing.

The technique we propose here, inspired by Perona and Malik's work on edge detection [Per90], can be applied to a depth map produced by any stereo matching algorithm. It utilizes neighboring information in a natural manner through nonlinear diffusion, filling large gaps while maintaining sharp object boundaries. The main difference between Perona and Malik's work and ours is that we use the gradient of the original image rather than that of the depth image. Hence, discontinuities in depth are assured to be consistent with intensity discontinuities, often a desirable property [Gam87]. Assuming that object shapes rarely vary dramatically apart from boundaries, this technique can eliminate many errors caused by textureless regions or occlusions.

The remainder of this paper is organized as follows. In Section 2, a generalized multi-baseline stereo is presented, which is used to produce all of our sample depth maps. Section 3 describes our method for improving depth maps by nonlinear diffusion, Section 4 discusses several applications that can benefit from such an improvement, and finally, some questions and directions for future research are discussed in Section 5.

2 GENERALIZED MULTIPLE-BASELINE STEREO

Okutomi and Kanade [Oku93] proposed a multiple-baseline stereo algorithm that applied to parallel cameras (i.e., there are only horizontal disparities). For a general camera setup, images must first be rectified, which usually requires a re-sampling of the images, during which some information may be lost. Here, we generalize the Okutomi and Kanade algorithm to deal with an arbitrary camera configuration.

First, we assume that all cameras are calibrated, for example, using Tsai's method [Tsa87]. The parameters of camera i are represented by $\mathbf{M_i}$. Knowing the center of projection for the camera, we may compute the ray $\mathbf{r}$ passing through the center and a given pixel $\mathbf{p} = (x, y)$ in the image. If one dimension is known of the real world point $\mathbf{P} = (X, Y, Z)$ corresponding to the pixel $\mathbf{p}$, for example, if we know $Z = c$, then we can compute the 3D position of $\mathbf{P}$ by intersection of ray $\mathbf{r}$ with the plane at $Z = c$, and from this, we may also compute its projection in the image planes of the other cameras, as shown in Fig. 1, where C_i, C_j are centers of projection for camera i, j, respectively.

Suppose $\mathbf{q} = \mathbf{q}(\mathbf{p}, Z, \mathbf{M_i}, \mathbf{M_j})$ is a function relating pixel $\mathbf{p}$ in camera i to pixel $\mathbf{q}$ in camera j. Given

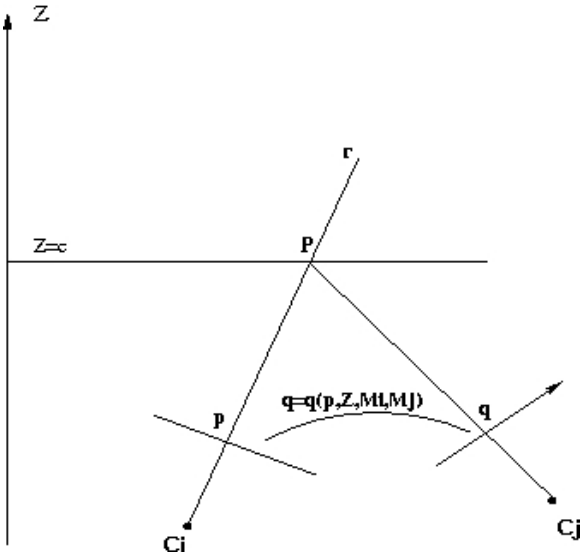


Figure 1: Scene point computing and reprojection

that the sum of squared differences (SSD) between two corresponding regions in an image pair can be used to measure similarity between these images, the sum of SSD (SSSD) over all image pairs may be used to determine the depth.

$$\mathrm{SSSD}(\mathbf{p}, Z) = \sum_{i \neq j} \sum_{\mathbf{p}' \in \mathbf{N_p}} (I_i(\mathbf{p}') - I_j(\mathbf{q}(\mathbf{p}', Z, \mathbf{M_i}, \mathbf{M_j})))^2 \quad (1)$$

where $\mathbf{N_p}$ is the neighborhood of pixel $\mathbf{p}$, $I_i(\mathbf{p})$ is the intensity of pixel $\mathbf{p}$ in camera i.

We take camera i as a reference and compute the sum of SSDs between the images obtained by camera i and all other cameras. The best depth estimate for each pixel $\mathbf{p}$ is the value of Z that minimizes the SSSD:

$$Z(\mathbf{p}) = argmin_Z \mathrm{SSSD}(\mathbf{p}, Z) \quad (2)$$

For best results, Z should be discretized as finely as possible subject to computational constraints.

Unfortunately, this approach yields unsatisfactory results, as illustrated in Fig. 2 with three cameras. The original images are pictured in the first column and their corresponding depth maps in the second. While the results are generally reasonable, there remain many errors, typically visible as bright points and black holes on the body, caused by occlusions, textureless regions, repeated patterns, and depth discontinuities.

3 POST-PROCESSING DEPTH MAPS BY NONLINEAR DIFFUSION

If two nearby pixels are in the same region or belong to the same object, their respective depths should

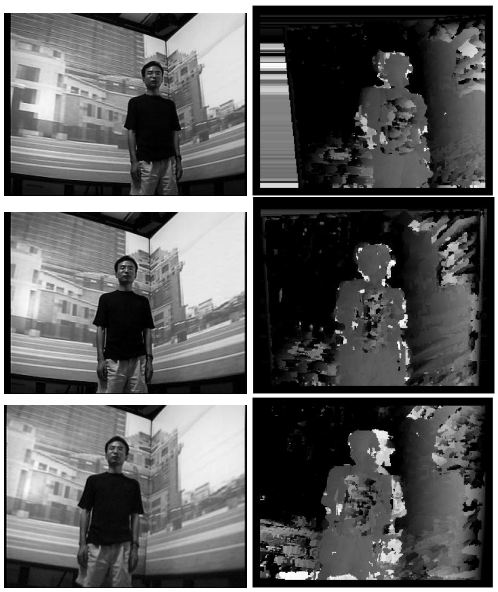


Figure 2: Original camera images (left column) and their corresponding depth maps (right column).

be similar. One way to achieve this smoothness constraint is to apply a weighted averaging, such as Gaussian smoothing, to the depth map. Unfortunately, such techniques also tend to blur boundaries, a problem we would like to avoid. Borrowing from Perona and Malik [Per90], who suggested an anisotropic diffusion method to improve edge detection, we apply the same technique to depth maps. Consider an updating function:

$$Z(x,y)_t = Z(x,y)_{t-1} + \lambda \mathbf{c}(x,y,t) \cdot \bigtriangledown \mathbf{Z} \quad (3)$$

where λ is a constant for numerical stability[1] and $\bigtriangledown \mathbf{Z}$ indicates nearest-neighbor differences. To achieve the desired effect, the coefficient $\mathbf{c}(x,y,t)$ should be high in the interior of each region, low at boundaries, and should have a steep threshold between the two cases. We note that the gradient G of the intensity image tends to large values along edges and small values in interior regions. Thus, an excellent choice for c is a function that responds maximally to low values of gradient, i.e. $c(x,y,t) = f(G)$, where $f(\cdot)$ takes on some shape approximating that shown in Fig. 3, for example,

$$f(G) = e^{-(\|G\|)/K)^2)} \quad (4)$$

or

$$f(G) = (1 + (\|G\|/K)^2)^{-1} \quad (5)$$

Eq. 5 is used here, which favors wide regions over smaller ones [Per90]. Unlike Perona and Malik's approach [Per90], we smooth the depth map based on

[1]For the results illustrated in this paper, we use a value of $\lambda = 0.25$

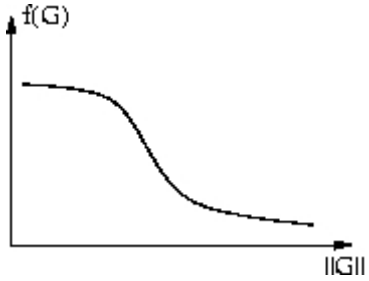


Figure 3: The qualitative shape of $f(\cdot)$.

the gradient of the original intensity image rather than that of the depth map itself. In this manner, we incorporate edge information into the depth map so as to recover the regions of occlusion. Through an iterative update as described by Equation 3, the depth map can be smoothed as desired, with significant improvements to the resulting depth map, as illustrated in Fig. 4. The depth errors, seen as holes in the subject's body and bright points near boundaries, are gradually smoothed out, while the boundaries are kept sharp. The main disadvantage of such an iterative method is computational cost. For this example of a 320x240 resolution image, our implementation required a total running time of approximately 3.6 seconds under Matlab on a Pentium III 1.1 Ghz machine. The iteration process stops either after a certain number of iterations or when the sum of depth changes over all pixels from one iteration to the next is below some predefined threshold.[2]

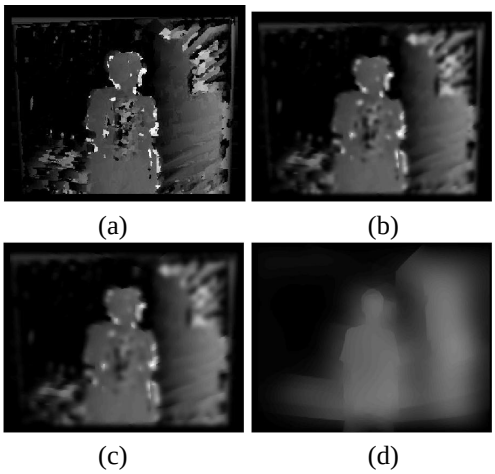


Figure 4: Results of nonlinear iteration diffusion: a)original depth map, b) after 10 iterations c) after 20 iterations, c) after 200 iterations.

[2]In the example shown here, the threshold was equivalent to a mean depth difference of 0.01.

4 APPLICATIONS

In this section, we summarize two important applications of our nonlinear diffusion technique for improved depth maps, namely, 3D scene reconstruction and background removal.

4.1 3D scene reconstruction

Starting from an intensity image and a corresponding depth map, it is possible to synthesize a novel view from a nearby viewing position, as illustrated in Fig. 5. Given the parameters of a virtual camera, the 3D position corresponding to a pixel can be computed from its position and depth. A surface can then be polygonized as described by Kanade *et al.* [Kan97]. A mesh (two triangles) is constructed, using the 3D coordinates of four neighboring pixels as the vertices of the triangles and the texture is obtained from the corresponding intensity map.

Due to estimation errors, discontinuities in the depth map, and the displacement between the virtual camera and the reference camera, some artificial surfaces typically appear in the synthesized view, as shown in Fig. 5a. These can be eliminated by adding smoothness constraints, that is, the mesh will not be rendered unless the depths of the three vertices of a triangle are similar. Unfortunately, this results in the appearance of *holes* in the image, as seen in Fig. 5b, 7d,g. However, using our improved depth map, as described above, the result appears to be improved greatly, as pictured in Fig. 5c,7e.

4.2 Background removal

In virtual reality and immersive telepresence applications, it is of critical importance to extract foreground objects (typically people) from the background. Since the background may change dynamically, it is often infeasible to perform such segmentation based on a 2D reference image, such as that employed by bluescreen techniques [Smi96]. Instead, we wish to perform this task based on 3D information from a CAVE-like environment. Captured images typically contain two perpendicular screens as background, which we can represent by the planes $X = 0$ and $Y = 0$. Since the environmental geometry is relatively simple, 3D scene information can be estimated from the depth map easily and we can then separate objects from the background by thresholding based on depth estimates. Sample results are illustrated in Fig. 6, 7f,g.

Due to the estimation errors of the original depth map, the segmented images in Fig. 6a and 7f include some portions of the background and some holes in the foreground. Using the improved depth map, instead, the results, pictured in Fig. 6b and 7g, are significantly improved, although still imperfect. The remaining problems are due to the fact that the diffusion effect is determined by image gradient information, which may not be consistent with the scene geometry. In Fig. 6 and 7, the background consists of planar screens, but the corresponding gradients are not flat because of the complex projected images appearing on them. We note that this situation is likely to pose problems for many stereo matching techniques, making use only of the visible light spectrum. As a result, occlusions still produce some artifacts near object boundaries, which cannot be removed entirely by diffusion. Due to the difficulty of tuning the diffusion process, smoothing over boundaries may still occur, thereby resulting in the occasional depth error.

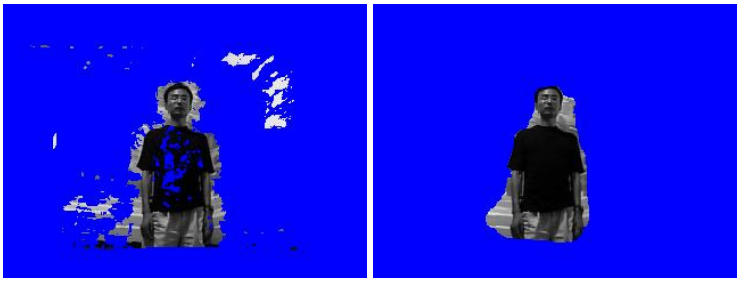
5 DISCUSSION

We have demonstrated that depth maps can be improved by nonlinear diffusion techniques, reducing the problems caused by textureless regions and occlusions. Since the post-processing step of the algorithm is based simply on image gradient information, the method may be applied to the benefit of a wide range of applications.

However, it is clear that nonlinear diffusion is not a panacea. The amount of improvement possible to the depth map is limited by the initial quality of the stereo matching algorithm; errors in the initial depth estimates tend to be propagated during the diffusion steps. Thus, it would be useful to have some method of evaluating the quality of the initial depth map. While no reliable measurements exist to evaluate the quality of stereo matching results, certain cues may be useful. For example, Scharstein and Szeliski [Sch96] proposed a disparity certainty measure for each location based on *winner margin* or *entropy*, which may be used to estimate an overall certainty. Such certainty measures can be used to determine automatically the need for a post-processing step and might be able to suggest an earlier stopping criteria for the nonlinear diffusion iteration. A similarity score (e.g. NCC) of a pixel can indicate the confidence of a match, i.e., the correct match should have a high score, although the converse is not necessarily the case. Egnal *et al* [Egn02] suggested a stereo confidence metric based on such cues. An improvement to our method may be obtained by weighting areas based on the degree of confidence in the corresponding area of the original depth map, thus reducing the influence of errors in the initial depths. Another issue affecting performance is the choice of a better *edge-stopping* function than the one currently used (Eq. 5). For example, Black *et al.* [Bla98] analyzed anisotropic diffusion in a statistical framework and related this technique to robust estimators and regularization with a line process. In our con-

(a) (b) (c)

Figure 5: A novel synthesized view a)based on the original depth map b) with the addition of smoothness constraints c) using the improved depth map of Fig. 4d, generated by nonlinear diffusion.



(a) (b)

Figure 6: Segmented images based on a) the original depth map, b) the improved depth map from Fig. 4d.

tinuing research, we hope to develop such ideas further.

6 ACKNOWLEDGMENTS

The completion of this research was made possible thanks to Bell Canada's support through its Bell University Laboratories R & D program. This support is gratefully acknowledged. The author would also like to thank Reg Wilson for sharing his implementation of Tsai's camera calibration method.

References

[Bel92] Belhumeur, P. and Mumford, D. A Bayesian Treatment of the stereo correspondence problem using half-occluded regions. Proc. CVPR'92, pp. 506-512, 1992.

[Bla98] Black, M. J., Sapiro, G., Marimont, D. H., and Heeger, D. Robust Anisotropic Diffusion. IEEE Transaction on Image Processing, Vol. 7, No. 3, pp. 421-432, 1998.

[Boy01] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast proximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 23, No. 11, pp. 1222-1239, 2001.

[Egn02] Egnal, G. Mintz, M., and Wildes, R. A stereo confidence metric using single view imagery. Proc. Vision Interface, pp. 162-170, 2002.

[Gam87] Gamble, E. and Poggio, T. Visual Integration and Detection of Discontinuities: The Key Role of Intensity Edges. MIT AI Lab Memo 970, 1987

[Gei92] Geiger, D., Ladendorf, B. and Yuille, A. Occlusion and binocular stereo. ECCV'92, pp. 425-433, 1992.

[Int94] Intille, S. S. and Bobick, A. F. Incorporating Intensity Edges in the Recovery of Occlusion Regions. International Conference on Pattern Recognition, pp. 674-677, 1994.

[Kan94] Kanade,T. and Okutomi, M. A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment. IEEE Trans. Pattern Recognition and Machine Intelligence, Vol. 16, No. 9, pp. 920-932, 1994.

[Kan97] Kanade, T., Rander, P., and Narayanan, P. J. Virtualized Reality: Constructing Virtual Worlds from Real Scenes. IEEE Multimedia, Immersive Telepresence, Vol. 4, No. 1, pp. 34-47, 1997.

[Kau01] Kauff, P., Brandenburg, N., Karl, M., and Schreer, O. Fast Hybrid Block- and Pixel- Recursive Disparity Analysis for Real-Time Applications in Immersive Tele-Conference Scenarios. Proc. of 9th International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision, pp. 198-205, 2001.

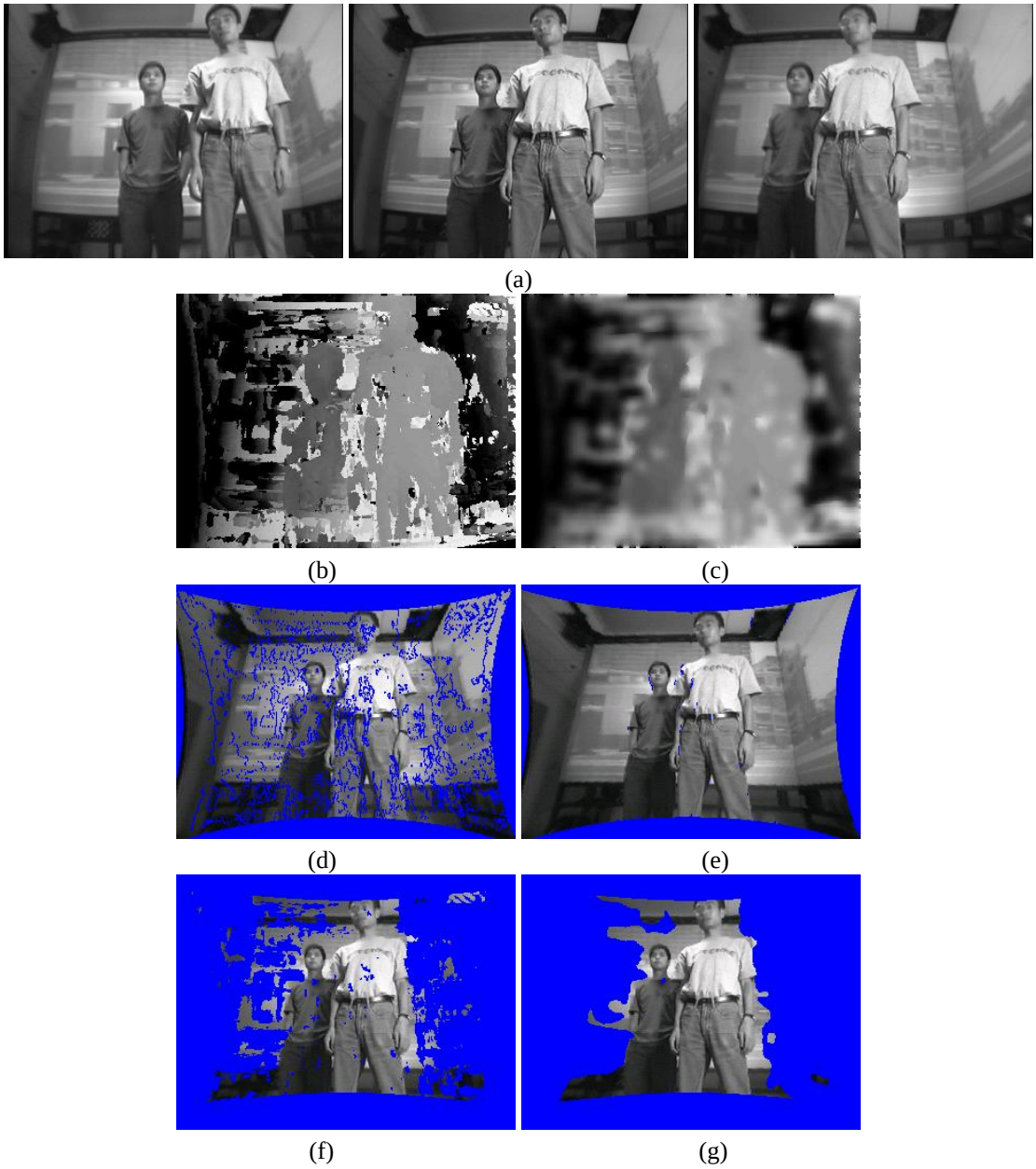


Figure 7: (a) original images, (b) depth map, (c) depth map after diffusion, (d) reconstruction based on (b), (e) reconstruction based on (c), (f) segmentation based on (b), (g) segmentation based on (c)

[Kol01] Vladimir Kolmogorov and Ramin Zabih, Visual correspondence with occlusions using graph cuts. International Conference on Computer Vision, pp. 508-515, 2001.

[Oku93] Okutomi, M. and Kanade, T. A Multiple-Baseline Stereo. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, pp. 353-363, 1993.

[Per90] Perona, P. and Malik, J. Scale-space and Edge Detection Using Anisotropic Diffusion. IEEE Trans. Pattern Recognition and Machine Intelligence, Vol. 12, No. 7, pp. 629-639, 1990.

[Roy98] Roy, S. and Cox, I. A maximum-flow formulation of the n-camera stereo correspondence problem. International Conference on Computer Vision, pp. 492–499, 1998.

[Sch96] Scharstein, D. and Szeliski, R. Stereo Matching with Non-linear Diffusion. CVPR'96, pp. 343-350, 1996.

[Sch02] Scharstein, D. and Szeliski, R. A Taxtonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV 47(1-3), pp. 7-42, 2002.

[Schm02] Schmidt, J., Niemann, H., and Vogt, S. Dense Disparity Maps in Real-Time with an Application to Augmented Reality. the Sixth IEEE Workshop on Applications of Computer Vision, pp. 225-230, 2002.

[Smi96] Smith, A. R. and Blinn, J. F. Blue Screen Matting. SIGGRAPH'96 Conference Proceedings, Annual Conference Series, pp. 259-268, 1996.

[Tsa87] Tsai, R. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses, IEEE J. Robotics and Automation, Vol. RA-3, No. 4, pp. 323-344, 1987.